NutritionVerse: Empirical Study of Various Dietary Intake Estimation Approaches

Chi-en Amy Tai, Matthew Keller, Saeejith Nair, Yuhao Chen*, Yifan Wu, Olivia Markham, Krish Parmar, Pengcheng Xi, Heather Keller, Sharon Kirkpatrick, Alexander Wong

University of Waterloo
Waterloo, Ontario, Canada

National Research Council Canada
Ottawa, Ontario, Canada

National Research Council Canada
Conseil national de recherches Canada

Introduction

- Accurate dietary intake estimation is critical for informing policies and programs to support healthy eating, as malnutrition has been directly linked to decreased quality of life

- Computer vision and machine learning have been used to automatically estimate dietary intake from food images

- Lack of high-quality comprehensive food image dataset with diverse viewpoints, modalities and food annotations

Existing Datasets

Work	Public	Data							Dietary Info				
		# Img	# Items	Real	Mixed	# Angles	Depth	Annotation Masks	CL	M	P	F	CB
[1]	✓	18	3	Y	N	1			✓				
[4]	✓	646	41	Y	Y	1			✓				
[23]	✓	50,374	201	Y	Y	1			✓				
[19]	✓	2,978	160	Y	N	2			✓	✓			
[34]	✓	5,006	555	Y	Y	4	✓		✓	✓	✓	✓	✓
[27]		3000	8	Y	Y	2		✓	✓	✓			

Table 1: Overview of existing dietary intake estimation datasets compared to ours where Mixed refers to whether multiple food item types are present in an image, and CL refers to calories, M to mass, P to protein, F to fat, and CB to carbohydrate.

[34] Q. Thames, et al.. 2021. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Nashville, 8903–8911

Existing Datasets

Work	Public	Data							Dietary Info				
		# Img	# Items	Real	Mixed	# Angles	Depth	Annotation Masks	CL	M	P	F	CB
[1]	✓	18	3	Y	N	1			✓				
[4]	✓	646	41	Y	Y	1			✓				
[23]	✓	50,374	201	Y	Y	1			✓				
[19]	✓	2,978	160	Y	N	2			✓	✓			
[34]	✓	5,006	555	Y	Y	4	✓		✓	✓	✓	✓	✓
[27]		3000	8	Y	Y	2		✓	✓	✓			

Table 1: Overview of existing dietary intake estimation datasets compared to ours where Mixed refers to whether multiple food item types are present in an image, and CL refers to calories, M to mass, P to protein, F to fat, and CB to carbohydrate.

[34] Q. Thames, et al.. 2021. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Nashville, 8903–8911

What if there is a perfect dataset?

- What is the best approach for dietary assessment?
- Does depth information improve model performance?
- What is the impact of using synthetic data?

Proposed Datasets

Work	Public	Data							Dietary Info				
		# Img	# Items	Real	Mixed	# Angles	Depth	Annotation Masks	CL	M	P	F	CB
[1]	✓	18	3	Y	N	1			✓				
[4]	✓	646	41	Y	Y	1			✓				
[23]	✓	50,374	201	Y	Y	1			✓				
[19]	✓	2,978	160	Y	N	2			✓	✓			
[34]	✓	5,006	555	Y	Y	4	✓		✓	✓	✓	✓	✓
[27]		3000	8	Y	Y	2		✓	✓	✓			
NV-Real	✓	889	45	Y	Y	4		✓	✓	✓	✓	✓	✓
NV-Synth	✓	84,984	45	N	Y	12	✓	✓	✓	✓	✓	✓	✓

Table 1: Overview of existing dietary intake estimation datasets compared to ours where Mixed refers to whether multiple food item types are present in an image, and CL refers to calories, M to mass, P to protein, F to fat, and CB to carbohydrate.

[34] Q. Thames, et al.. 2021. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Nashville, 8903–8911

NV-Synth Dataset

- Using 3D meshes from the NutritionVerse-3D dataset and Nvidia's Omniverse IsaacSim simulation framework
- NV-Synth is a collection of 84,984 2D images of 7,082 distinct dishes with associated dietary metadata and labels

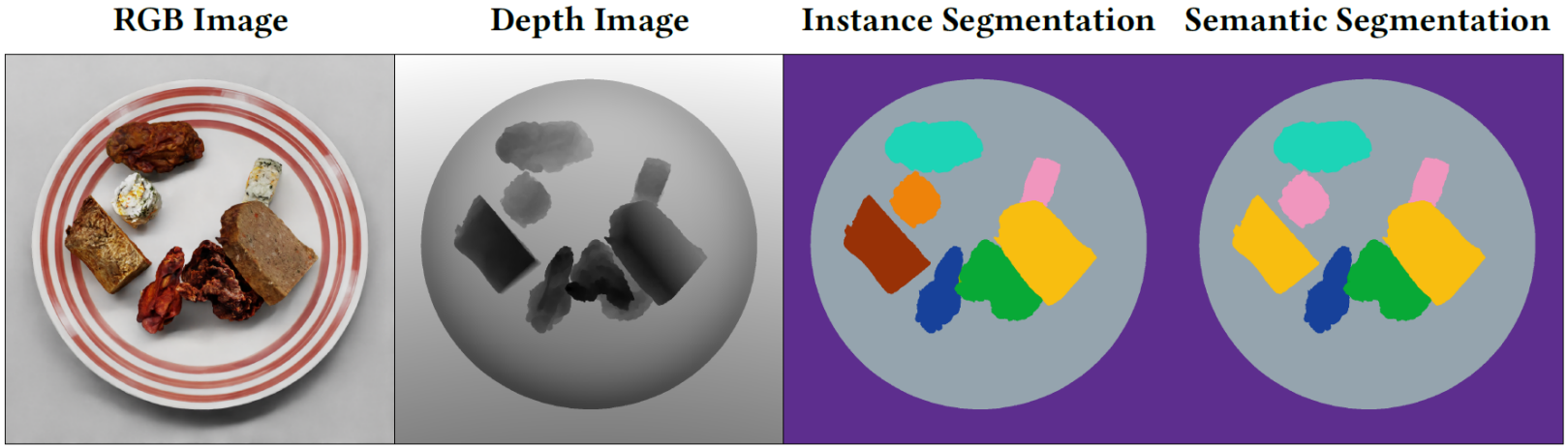


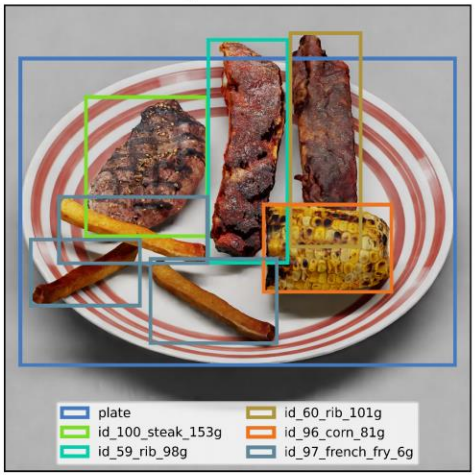
Figure 1: Sample scene from NV-Synth dataset with the associated multi-modal image data (e.g., RGB and depth data) and annotation metadata (e.g., instance and semantic segmentation masks) derived using objects from the NutritionVerse-3D dataset [33]. There are 2 meatloaves, 1 chicken leg, 1 chicken wing, 1 pork rib, and 2 sushi rolls in this scene.

Synthetic Data Generation

- Each sample contains
  - Nutrition information
  - Segmentation masks
    - Semantic segmentation
    - Instance segmentation
    - Amodal instance segmentation
  - RGB and Depth

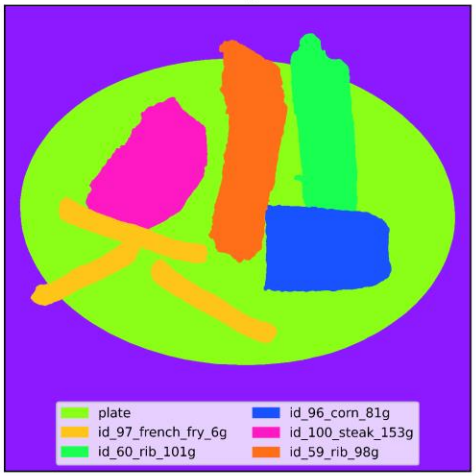
RGB Image

2D Object Detections



Depth Image

Semantic Segmentation



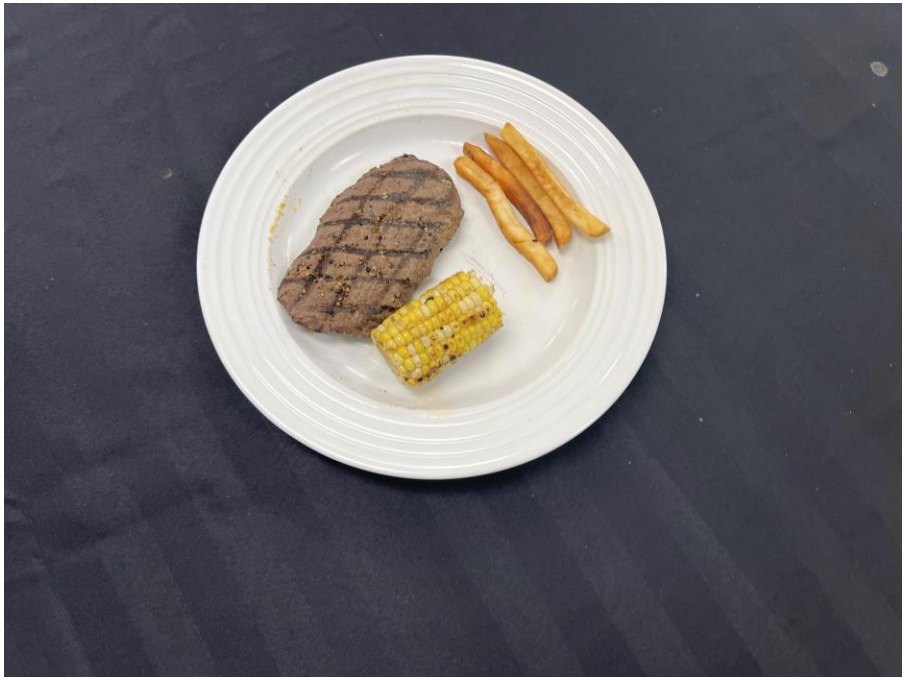
NV-Synth Dataset

- To generate synthetic scenes of meals, up to 7 ingredients were sampled and then procedurally dropped onto a plate to simulate realistic food scenes
- We filter all scenes with food dropped outside of the plate

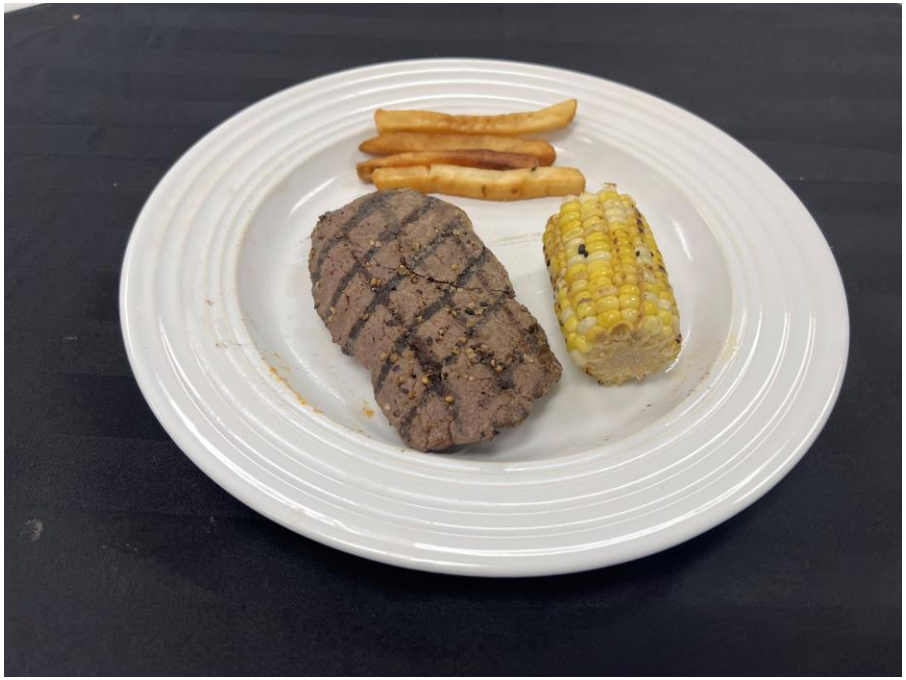
- Amodal segmentation (complete food segmentation mask even the food is occluded) is not provided by the Omniverse simulation software
  - We obtain the mask by iteratively turn on/off the visibility of other food items

NV-Real Dataset

- 889 Images, 251 distinct dishes, 45 food items
- Human labeled segmentation masks
- Nutrition information is recorded per-item instead of per-plate



View 1

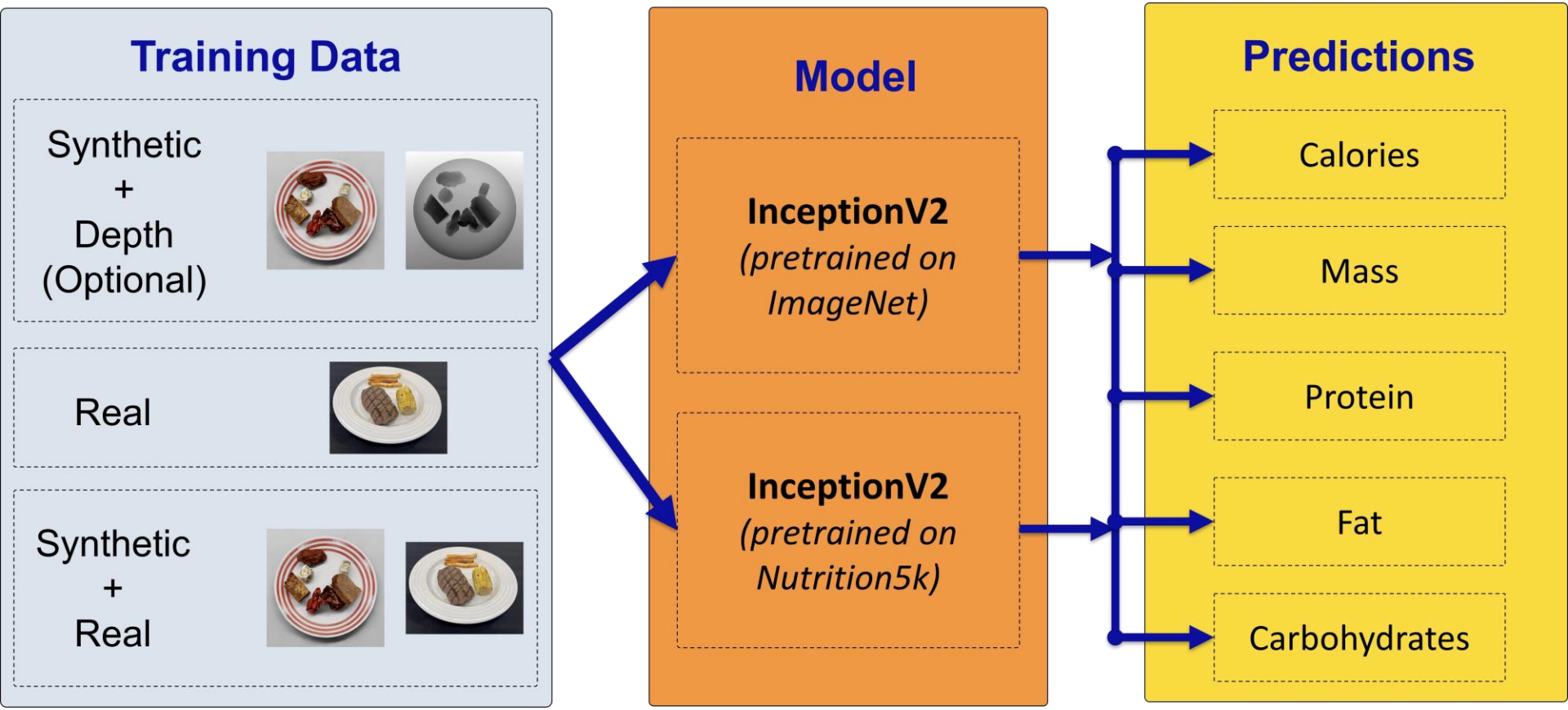


View 2

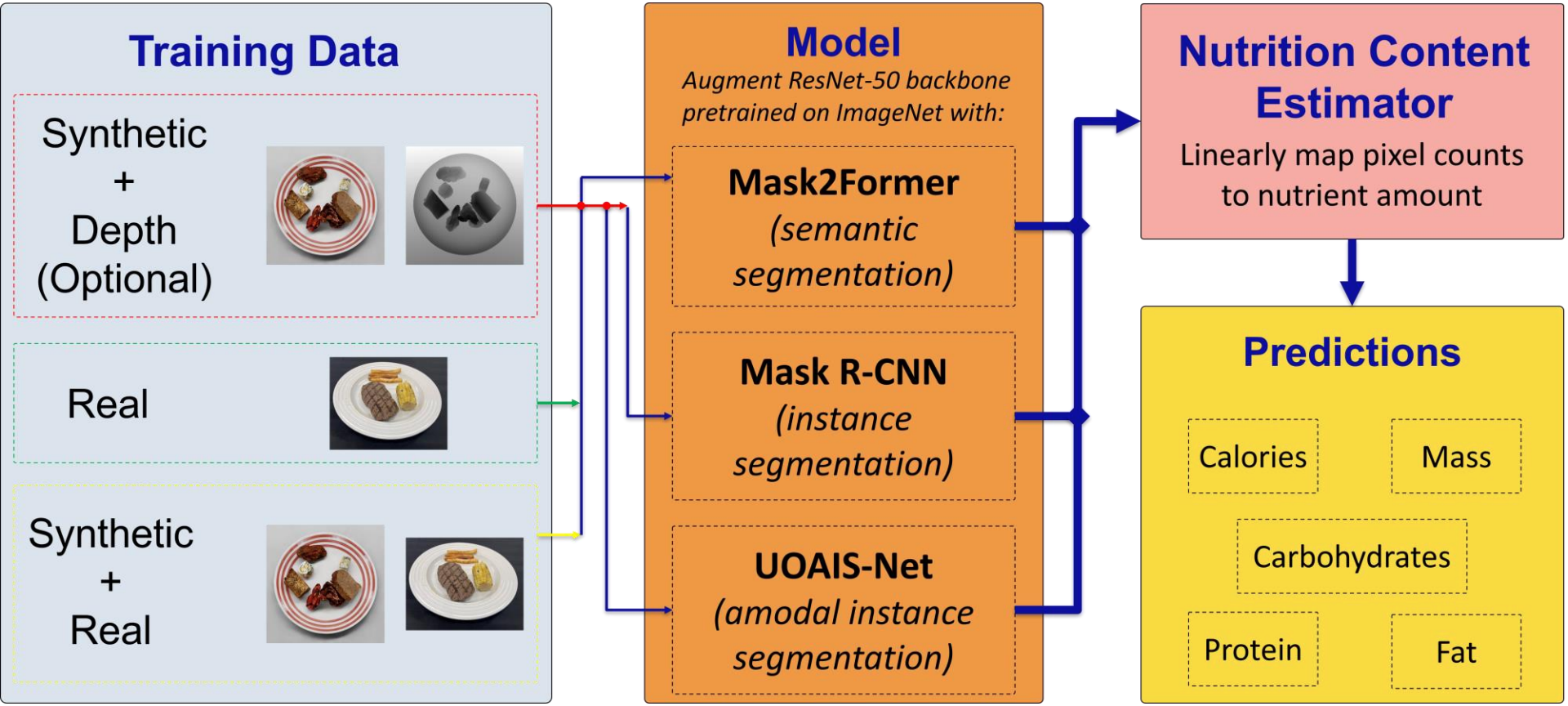
Experiment Setup

- Direct approach
  - InceptionV2 backbone (used in Nutrition5K [34])
  - Pretrained on ImageNet
  - Pretrained on Nutrition5K
- Indirect approach
  - Assume a fixed linear relationship between the pixel count of the food item and its nutrition value
  - Semantic segmentation
  - Instance segmentation
  - Amodal instance segmentation

Direct Approach



Indirect Approach



Example Results

(a) RGB input (Ground Truth)
CL: 1609, M: 684, P: 65, F: 69, CB: 183

(b) Semantic segmentation
CL: 371, M: 156, P: 21, F: 18, CB: 32

(c) Instance segmentation
CL: 706, M: 268, P: 39, F: 33, CB: 65

(d) Amodal instance segmentation
CL: 898, M: 337, P: 45, F: 40, CB: 90

Figure 8: Segmentation and prediction results of models trained with RGB input where CL refers to calories, M to mass, P to protein, F to fat, and CB to carbohydrate.

What is the best approach for dietary assessment?

- When given perfect labels from simulation, direct prediction gives the best nutrition estimation

Model (RGB)	Eval Dataset	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
Semantic	NV-Synth	418.1	185.4	39.0	23.5	32.3
Instance	NV-Synth	430.9	191.4	39.3	24.1	34.4
Amodal Instance	NV-Synth	451.3	202.8	39.6	24.8	38.5
Direct Prediction (ImageNet)	NV-Synth	229.2	102.6	56.0	12.0	**19.4***
Direct Prediction (Nutrition5k)	NV-Synth	**128.7***	**77.2***	**18.5***	**9.1***	21.5

Table 3: Evaluation of model architectures using NV-Synth (RGB images) with the lowest MAE value for each column bolded with an * next to it.

Does depth information improve model performance?

- The addition of depth information results in a worse performance for direct prediction models
- For indirect methods, depth does not make any significant difference

Model (RGB)	Eval Dataset	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
Semantic	NV-Synth	418.1	185.4	39.0	23.5	32.3
Instance	NV-Synth	430.9	191.4	39.3	24.1	34.4
Amodal Instance	NV-Synth	451.3	202.8	39.6	24.8	38.5
Direct Prediction (ImageNet)	NV-Synth	229.2	102.6	56.0	12.0	**19.4***
Direct Prediction (Nutrition5k)	NV-Synth	**128.7***	**77.2***	**18.5***	**9.1***	21.5

Table 3: Evaluation of model architectures using NV-Synth (RGB images) with the lowest MAE value for each column bolded with an * next to it.

Model (RGBD)	Eval Dataset	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
Semantic	NV-Synth	418.3	185.3	39.0	23.5	32.2
Instance	NV-Synth	432.9	194.1	39.1	24.1	35.2
Amodal Instance	NV-Synth	462.0	208.1	39.7	25.2	40.4
Direct Prediction (ImageNet)	NV-Synth	371.7	317.6	34.8	**19.2***	**25.2***
Direct Prediction (Nutrition5k)	NV-Synth	**202.0***	**78.8***	**23.5***	30.1	33.3

Table 4: Investigation of depth information using NV-Synth (RGBD images) with the lowest MAE value for each column bolded with an * next to it.

What is the impact of using synthetic data?

- We investigate this question by comparing the for three scenarios:

(A) Using models trained only on NV-Synth

(B) Fine-tuning models trained on NV-Synth using NV-Real

(C) Training models only on NV-Real

Using models trained only on NV-Synth

- With perfect data, direct prediction pre-trained on Nutrition5k provides the best performance

Model (Scenario A)	Eval Dataset	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
Semantic	NV-Real	40830.5	17342.0	2086.4	1630.4	4432.3
Instance	NV-Real	50190.0	33774.6	2950.5	2009.5	5108.0
Amodal Instance	NV-Real	72999.6	38379.2	4460.2	3225.3	6580.1
Direct Prediction (ImageNet)	NV-Real	530.6	**182.9***	62.6	27.7	**54.4***
Direct Prediction (Nutrition5k)	NV-Real	**525.9***	188.4	**39.1***	**27.4***	54.6

Table 5: Scenario A: Models trained only on NV-Synth, with the lowest MAE value for each column bolded with an * next to it.

Fine-tuning models trained on NV-Synth using NV-Real

- We train models with our large-scale NV-Synth dataset and fine-tune the trained models with NV-Real dataset to close the sim-to-real gap
- Direct prediction with ImageNet pretrained model provides the best performance
  - This highlights the high domain-generalization capability of the ImageNet pretrained features

Model (Scenario B)	Eval Dataset	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
Semantic	NV-Real	445.1	219.2	39.0	22.8	**41.8***
Direct Prediction (ImageNet)	NV-Real	**229.8***	**63.3***	**24.6***	**13.5***	70.8
Direct Prediction (Nutrition5k)	NV-Real	813.3	407.5	59.6	36.1	68.8

Table 6: Scenario B: Models trained on NV-Synth and fine-tuned on NV-Real, with the lowest MAE value for each column bolded with an * next to it.

Training models only on NV-Real

- For sanity check, we train and evaluate models only on the NV-Real dataset
- Semantic segmentation model provides the best result
  - When the dataset is small, prior knowledge helps

Model (Scenario C)	Eval Dataset	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
Semantic	NV-Real	**442.7***	**221.0***	**40.1***	**23.0***	**42.4***
Direct Prediction (ImageNet)	NV-Real	778.8	386.6	56.8	37.6	65.9
Direct Prediction (Nutrition5k)	NV-Real	624.9	426.5	46.5	37.7	70.8

Table 7: Scenario C: Models trained only on NV-Real, with the lowest MAE value for each column bolded with an * next to it.

What is the impact of using synthetic data?

- We investigate this question by comparing the for three scenarios:

(A) Using models trained only on NV-Synth

(B) Fine-tuning models trained on NV-Synth using NV-Real

(C) Training models only on NV-Real

Model Description	Trained	Fine-Tuned	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
(A) Direct Prediction (Nutrition5k)	NV-Synth	N/A	525.9	188.4	39.1	27.4	54.6
(B) Direct Prediction (ImageNet)	NV-Synth	NV-Real	**229.8***	**63.3***	**24.6***	**13.5***	70.8
(C) Semantic	NV-Real	N/A	442.7	221.0	40.1	23.0	**42.4***

Table 8: Comparison of the best model from the three scenarios evaluated on the NV-Real dataset, with the lowest MAE value for each column bolded with an * next to it.

- The best model is **Direct Prediction (ImageNet) model trained on the NV-Synth train set and fine-tuned on the NV-Real train set**

Conclusion

- Introduce two new food datasets
  - NV-Synth (simulation)
  - NV-Real (manually collected)
- We investigated various intake estimation approaches and concluded that
  - Direct prediction produces the best result
  - Depth does not help much and may make results worse
  - Synthetic dataset boosts performance

UNIVERSITY OF WATERLOO

RIA Research Institute for Aging
Schlegel • UWaterloo • Conestoga
Enhancing Life

VIP lab

National Research Council Canada
Conseil national de recherches Canada

Thank You!

NV-Real dataset
NV-Synth dataset

NutritionVerse-3D